Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Examples and numerics

Paul Dupuis* and Hui Wang†
Lefschetz Center for Dynamical Systems
Brown University
Providence, R.I. 02912
USA

August 8, 2005

Abstract

It was established in [4, 5] that importance sampling algorithms for estimating rare-event probabilities are intimately connected with two-person zero-sum differential games and the associated Isaacs equation. The purpose of the present paper and a companion paper [6] is to show that the classical sense *subsolutions* of the Isaacs equation can be used as a basic and flexible tool for the construction and analysis of efficient importance sampling schemes. The importance sampling algorithms based on subsolutions are *dynamic* in the sense that during the course of a single simulation, the change of measure used at each time step may depend on the outcome of the simulation up until that time. While [6] focuses on a theoretical aspects, the present paper discusses explicit methods of constructing subsolutions, implementation issues, and simulation results.

*Research of this author supported in part by the National Science Foundation (NSF-DMS-0306070 and NSF-DMS-0404806) and the Army Research Office (DAAD19-02-1-0425).

†Research of this author supported in part by the National Science Foundation (NSF-DMS-0103669 and NSF-DMS-0404806).

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
08 AUG 2005		**00-08-2005 to 00-08-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Examples and numerics	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Brown University,Division of Applied Mathematics,182 George Street,Providence,RI,02912**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**43**	

1 Introduction and background

It was established in [4, 5] that importance sampling algorithms for estimating rare-event probabilities or functionals that are largely determined by rare-events are closely related to deterministic differential games. More precisely, the asymptotic optimal performance of importance sampling schemes can be characterized by the value function of a two-person zero-sum differential game, which can in turn be characterized by the solution to the Isaacs equation (a nonlinear PDE) associated with the game. It was also discussed in [4, 5] that one can construct asymptotically optimal importance sampling algorithms based on this solution.

The purpose of the present paper and a companion paper [6] is to explore this connection in further depth. A main new feature is that it is possible to construct efficient importance sampling schemes based on *subsolutions* of the Isaacs equation. This leads to a more general class of schemes and lends great flexibility to the designer. We will see that one can often construct subsolutions that are structurally simpler than the actual *solution*, but which correspond to asymptotically optimal (or at least nearly asymptotically optimal) importance sampling schemes that reflect this simpler structure. This simplicity is important, since one is often interested in properties other than just asymptotic optimality, e.g., ease of construction and ease of implementation. The companion paper [6] focuses on theoretical aspects of this approach, and proves a basic result on the asymptotic performance of importance sampling schemes that are based on subsolutions. In contrast, the present paper will focus on methods for constructing subsolutions and their application to various process models, events, and expected values.

The theory in [6] assumes a setting that includes as special cases both the sum of independently identically distributed (iid) random variables and the empirical measure of a finite-state Markov chain. However, its potential application is much broader, and includes, for example, systems with state dependencies, systems with constrained dynamics, and expected values involving path-dependent events. Indeed, while many examples discussed in the present paper do not fit directly into the theoretical framework of [6], we will see that efficient importance sampling algorithms can be designed via subsolutions.

The paper is organized as follows. In Section 2 we focus on a particular problem of interest: importance sampling for sums of a functional of a Markov chain. This is a special case of the model studied in [6]. We first recall the Isaacs equation that is appropriate for this problem, and then review

the notion of a *subsolution/control pair* for this equation. Such pairs define importance sampling schemes, and we also review the theoretical bounds on performance one obtains from a subsolution. Section 3 discusses methods for the construction of subsolutions for this particular problem in great detail, starting with the simplest possible examples and then extending to more complicated situations. Section 4 is devoted to examples. Whenever possible, we have chosen examples for which there is an alternative method to compute (at least approximately) the true value. The only exception is the example of Section 4.5. The first part of Section 4 the applies the techniques of Section 3 to problems involving sums of random variables, and presents simulation results for the corresponding schemes. However, the use of subsolution methods extends far beyond the simple setting of sums of random variables. For example, the paper [3] shows how to apply subsolutions to construct and analyze importance sampling schemes for stochastic networks. The second part of Section 4 considers several other types of problems to which the approach can be applied, including multi-dimensional level crossing problems, probabilities and expected values that depend on a sample path of a process, and an interesting mixed open/closed queueing network. Although the theoretical analysis of these problems is not covered by [6], in each case the application of the Isaacs equation and subsolution approach follows the pattern laid out for the simple case of sums of random variables. The examples are representative, but far from exhaustive. To streamline the presentation the details of certain constructions are postponed to an appendix.

2 Review of the main theoretical results

This section reviews the theoretical results in [6]. To ease exposition, we specialize to Markov chains and leave out various technical assumptions on the underlying processes and relevant functionals. See [6] for these details.

2.1 Basic setup

Let $Y \doteq \{Y_i, i \in \mathbb{N}_0\}$ denote a Markov chain with state space S. Assume that S is a Polish space, and let $p(y, dz)$ denote the probability transition kernel. Our interest is in sums of the form

$$X_n \doteq \frac{1}{n} \sum_{i=0}^{n-1} g\left(Y_i\right),$$

where $g : S \to \mathbb{R}^d$ is a given measurable function.

An important component of the analysis is the eigenvalue/eigenvector problem associated with the non-negative kernel $\exp\langle\alpha, g(z)\rangle p(y, dz)$ for an arbitrary $\alpha \in \mathbb{R}^d$:

$$\int_S \int_{\mathbb{R}^d} e^{\langle\alpha, g(z)\rangle} r(z;\alpha) p(y, dz) = e^{H(\alpha)} r(y;\alpha). \tag{2.1}$$

It follows from (2.1) that, for each $\alpha \in \mathbb{R}^d$,

$$P(y, dz;\alpha) \doteq e^{\langle\alpha, g(z)\rangle - H(\alpha)} \frac{r(z;\alpha)}{r(y;\alpha)} p(y, dz) \tag{2.2}$$

defines a probability transition kernel. The large deviation rate function L associated with the process $\{X_n\}$ can be expressed in terms of the eigenvalue H. Indeed, L is the convex conjugate of H, that is, for each $\beta \in \mathbb{R}^d$,

$$L(\beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle\alpha, \beta\rangle - H(\alpha)].$$

2.2 The Isaacs equation and subsolution

Suppose one wishes to estimate the expected value of certain functionals of X_n for large n, using importance sampling algorithms based on changes of measure of the form (2.2). It follows from [5] that the optimal performance of such schemes can be characterized by the Isaacs equation

$$W_t + \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(DW; \alpha, \beta) = 0 \tag{2.3}$$

with a suitable terminal condition. Here $W : (x, t) \in \mathbb{R}^d \times [0, 1] \to \mathbb{R}$, W_t denotes the partial derivative with respect to t, DW the gradient in x, and

$$\mathbb{H}(s;\alpha,\beta) \doteq \langle s, \beta\rangle + L(\beta) + \langle\alpha, \beta\rangle - H(\alpha) \tag{2.4}$$

for $s, \alpha, \beta \in \mathbb{R}^d$.

In order to construct good importance sampling schemes, one does not need the solution to the Isaacs equation. It turns out that finding a good subsolution (more precisely, a subsolution/control pair) is often sufficient. We need the following definition. Given $K \in \mathbb{N}$, consider a function $\bar{W} : \mathbb{R}^d \times [0, 1] \to \mathbb{R}$ and, for $1 \le k \le K$, functions $\rho_k : \mathbb{R}^d \times [0, 1] \to \mathbb{R}$, $\bar{\alpha}_k : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$.

Definition 2.1 *The collection* $(\bar{W}, \rho_k, \bar{\alpha}_k)$ *will be called a* generalized subsolution/control *if the following conditions hold.* $\rho_k : \mathbb{R}^d \times [0,1] \to \mathbb{R}, k = 1, \ldots, K$ *is a partition of unity, i.e., each* ρ_k *is non-negative, and*

$$\sum_{k=1}^{K} \rho_k(x,t) = 1$$

for all $(x,t) \in \mathbb{R}^d \times [0,1]$. $\bar{W}_t$ *and* $D\bar{W}$ *have representations*

$$\bar{W}_t(x,t) = \sum_{k=1}^{K} \rho_k(x,t) r_k(x,t), \quad D\bar{W}(x,t) = \sum_{k=1}^{K} \rho_k(x,t) s_k(x,t),$$

and for each $k = 1, \ldots, K$

$$r_k(x,t) + \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(s_k(x,t); \bar{\alpha}_k(x,t), \beta) \geq 0.$$

The functions $(r_k, s_k, \rho_k, \bar{\alpha}_k)$ *are uniformly bounded and Lipschitz continuous.*

For any generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$, it is not difficult to show that

$$\bar{W}_t + \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}; \alpha, \beta) \geq 0.$$

In other words, $\bar{W}$ is a classical subsolution to the Isaacs equation (2.3). It will turn out that the $(\rho_k, \bar{\alpha}_k)$-component determines the change of measure used in importance sampling, and so we use terminology "subsolution/control."

Remark 2.1 For the special case where $K = 1$, and with notation $\bar{\alpha} = \bar{\alpha}_1$, we simply write $(\bar{W}, \bar{\alpha})$ and call it a *subsolution/control* pair.

Remark 2.2 Suppose that $\bar{W}$ is a classical subsolution to the Isaacs equation (2.3), that is

$$\bar{W}_t + \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}; \alpha, \beta) \geq 0.$$

Let $\alpha^*(x,t)$ be the supremizer in the above display [indeed, one can easily identify $\alpha^*(x,t) = -D\bar{W}(x,t)/2$]. Then $(\bar{W}, \alpha^*)$ is a subsolution/control pair, provided that α^* is uniformly bounded and Lipschitz continuous.

2.3 Importance sampling algorithms based on subsolutions

In this section we describe the importance sampling algorithms associated with a given generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$. For fixed n, let

$$\bar{\alpha}_{k,j}(x) \doteq \bar{\alpha}_k(x, j/n), \quad \rho_{k,j}(x) \doteq \rho_k(x, j/n).$$

Processes $\{\bar{X}_j\}$ and $\{\bar{Y}_j\}$ are constructed recursively as follows. Let $\bar{X}_0 = 0$ and $\bar{Y}_0 = Y_0$. Given that $\bar{X}_j = x$ and $\bar{Y}_j = y$, simulate $\bar{Y}_{j+1}$ under the distribution

$$\sum_{k=1}^{K} \rho_{k,j}(x) P\left(y, dz; \bar{\alpha}_{k,j}(x)\right). \tag{2.5}$$

In other words, one first simulates a multinomial random variable I taking values in $\{1, 2, \ldots, K\}$ such that $P\{I = k\} = \rho_{k,j}(x)$, and then simulates $\bar{Y}_{j+1}$ from the distribution $P(y, dz; \bar{\alpha}_{I,j}(x))$. Finally, update the state dynamics by letting

$$\bar{X}_{j+1} \doteq \bar{X}_j + \frac{1}{n} g(\bar{Y}_{j+1}).$$

An unbiased importance sampling estimator can then be obtained by multiplying the functional of interest by a Radon-Nikodým derivative (i.e., likelihood ratio). For example, suppose we are interested in estimating $E[G(X_n)]$ for some function G. A single sample of the unbiased importance sampling estimator is

$$\begin{aligned} Z^n \doteq G(\bar{X}_n) \prod_{j=0}^{n-1} \Bigg[\sum_{k=1}^{K} \rho_{k,j}(\bar{X}_j) \cdot e^{\left\langle \bar{\alpha}_{k,j}(\bar{X}_j), g(\bar{Y}_{j+1}) \right\rangle - H\left(\bar{\alpha}_{k,j}(\bar{X}_j)\right)} \\ \cdot \frac{r(\bar{Y}_{j+1}; \bar{\alpha}_{k,j}(\bar{X}_j))}{r(\bar{Y}_j; \bar{\alpha}_{k,j}(\bar{X}_j))} \Bigg]^{-1}. \end{aligned} \tag{2.6}$$

The importance sampling algorithm is then just the sample average of a number of independent replications of Z^n.

Remark 2.3 In most of the applications considered in this paper, one can construct a generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$ where the $\bar{\alpha}_k$ are all constants and with K of moderate size. This has a distinct advantage in numerical implementation. For example, in the Markov chain case, to compute a change of measure one often needs to numerically solve the eigenvalue/eigenvector problem. If $\bar{\alpha}_k$ is not a constant, one needs to solve eigenvalue/eigenvector problems over and over again, depending on the current state of the simulation. This could become very computationally demanding.

2.4 The main theoretical result

To simplify the exposition, we specialize for now to the case where the quantity of interest is

$$E[G(X_n)] = E[1_{\{X_n \in A\}}] = P\{X_n \in A\}.$$

for some Borel set $A \subset \mathbb{R}^d$, and assume the large deviation limit

$$\lim_{n\to\infty} \frac{1}{n} \log P\{X_n \in A\} = -\inf_{\beta \in A} L(\beta).$$

For convenience let $\gamma \doteq \inf_{\beta \in A} L(\beta)$.

The following result is a special case of Theorem 7.1 of [6].

Theorem 2.2 *Let $(\bar{W}, \rho_k, \bar{\alpha}_k)$ be a generalized subsolution/control such that*

$$\bar{W}(x, 1) \leq 0 \tag{2.7}$$

for all $x \in A$. Let V^n be the second moment of the associated importance sampling estimator, that is,

$$V^n \doteq E\left[(Z^n)^2\right],$$

where Z^n is defined in (2.6). Let

$$W^n \doteq -\frac{1}{n} \log V^n.$$

Then

$$\liminf_{n\to\infty} W^n \geq \bar{W}(0, 0).$$

The theorem provides a lower bound for the exponential decay rate of the second moment of the importance sampling estimator Z^n. An upper bound that is independent of the importance sampling scheme also holds. Indeed, consider any importance sampling estimator one might use to approximate $P\{X_n \in A\}$, and analogous to W^n, let $\hat{W}^n$ denote the normalized logarithm of its second moment. Then an elementary calculation based on Jensen's inequality shows that

$$\limsup_{n\to\infty} \hat{W}^n \leq 2\gamma.$$

Thus a natural goal in importance sampling design is to come as close to this upper bound as possible. In other words, we would like to find a subsolution/control for which $\bar{W}(0, 0)$ is close to 2γ. Any importance sampling estimator that achieves this maximal exponential decay rate is called *asymptotically optimal*.

Remark 2.4 Theorem 2.2 holds under much greater generality [6]. For example, suppose one is interested in the expectation

$$E[G(X_n)] = E[\exp\{-nF(X_n)\}]$$

for some function F. A result analogous to Theorem 2.2 holds except that the terminal condition as in equation (2.7) is replaced by

$$\bar{W}(x,1) \leq 2F(x)$$

for all $x \in \mathbb{R}^d$.

2.5 Discussion

The use of subsolutions is applicable in much broader settings than those discussed in this section, or even those in [6]. For instance, an interesting class of problems is to estimate the probabilities of path dependent events (say, events that depend on the extrema of the sample path). In this case, the subsolution needs to satisfy certain boundary conditions besides the terminal condition. Another interesting class of problems where the use of subsolutions appears crucial involve probabilities associated with systems with constrained dynamics (e.g., queuing networks [3]), where the boundary condition and/or terminal condition may take more complicated forms. Depending on the class of changes of measure used and the dynamics of the system, the Isaacs equation may take a different form from the one used here. However, it is always the case that the use of subsolutions is critical to the construction of importance sampling schemes. Section 4 presents a few such examples even though they are not covered by the theoretical framework of [6].

3 Construction of generalized subsolution/controls

To illustrate how one can construct generalized subsolution/controls, we specialize to the setup of Section 2 and assume that the quantity of interest is

$$E[G(X_n)] = E[1_{\{X_n \in A\}}] = P\{X_n \in A\}.$$

The extrapolation to more general setups will also be considered later in the paper.

There are several concerns in the construction. To begin, the terminal condition $\bar{W}(x,1) \leq 0$ for $x \in A$ must be satisfied in order for Theorem

2.2 to be valid. Secondly, one wishes $\bar{W}(0,0)$ to be as close as possible to the optimal decay rate 2γ. Finally, one would like the controls $(\rho_k, \bar{\alpha}_k)$ to take simple forms, since this gives importance sampling algorithms that are simpler and easier to implement.

Our construction is as follows. We first identify a family of particularly simple subsolution/control pairs to the Isaacs equation (2.3). For each of these pairs, say $(\bar{W}, \bar{\alpha})$, $\bar{W}$ is affine in (x,t) and $\bar{\alpha}$ takes constant value. $\bar{W}$ will actually be a *solution* to the Isaacs equation alone (i.e., with the terminal condition unspecified). This family, denoted from now on by $\mathbb{A}$, will contain the building blocks for our construction. It is the appropriate family for the class of problems under consideration, where the Hamiltonian $\mathbb{H}$ does not depend on x.

If the terminal condition (2.7) is satisfied by an element of $\mathbb{A}$ and if $\bar{W}(0,0)$ is sufficiently close to 2γ, then the construction is complete. This is always possible when A is convex (see Case 1 in Section 3.2.1 below). When this is not possible, it is often possible to take a finite collection of such pairs, say $\{(\bar{W}_k, \bar{\alpha}_k), k = 1, 2, \ldots, K\}$, so that the minimum of $\{\bar{W}_k\}$ satisfies the terminal condition (2.7) and $\wedge_{k=1}^{K} \bar{W}_k(0,0)$ equals 2γ.

Since the each $\bar{W}_k$ is a classical subsolution, their minimum is a weak sense subsolution, but not a classical sense subsolution. However, there are several simple and easily implemented methods for mollifying $\wedge_{k=1}^{K} \bar{W}_k$ to produce a generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$. These are discussed in Section 3.3.

3.1 Affine solutions to the Isaacs equation

In this section we identify $\mathbb{A}$, the collection of affine subsolution/control pairs to the Isaacs equation (2.3).

For any given $\bar{\alpha} \in \mathbb{R}^d$ and $\bar{c} \in \mathbb{R}$, consider the affine function

$$\bar{W}(x,t) = -2\langle \bar{\alpha}, x \rangle + \bar{c} - 2(1-t)H(\bar{\alpha}).$$

We claim that $(\bar{W}, \bar{\alpha})$ is a subsolution/control pair. Indeed, thanks to the

convex conjugacy between H and L,

$$\begin{aligned}
&\bar{W}_t + \inf_{\beta\in\mathbb{R}^d} \mathbb{H}(D\bar{W};\bar{\alpha},\beta) \\
&= \bar{W}_t + \inf_{\beta\in\mathbb{R}^d} \left[\left\langle D\bar{W},\beta\right\rangle + L(\beta) + \langle\bar{\alpha},\beta\rangle - H(\bar{\alpha})\right] \\
&= 2H(\bar{\alpha}) + \inf_{\beta\in\mathbb{R}^d} \left[\langle -2\bar{\alpha},\beta\rangle + L(\beta) + \langle\bar{\alpha},\beta\rangle - H(\bar{\alpha})\right] \\
&= H(\bar{\alpha}) + \inf_{\beta\in\mathbb{R}^d} \left[L(\beta) - \langle\bar{\alpha},\beta\rangle\right] \\
&= 0.
\end{aligned}$$

Let $\mathbb{A}$ be the collection of all such functions, that is

$$\mathbb{A} \doteq \left\{(\bar{W},\bar{\alpha}) : \bar{W} = -2\langle\bar{\alpha},x\rangle + \bar{c} - 2(1-t)H(\bar{\alpha}), \bar{\alpha}\in\mathbb{R}^d, \bar{c}\in\mathbb{R}\right\}.$$

Remark 3.1 It is not difficult to show that for every $(\bar{W},\bar{\alpha})\in\mathbb{A}$, the affine function $\bar{W}$ is indeed a *solution* to the Isaacs equation (2.3).

3.2 Piecewise affine viscosity subsolutions

As mentioned above, the technique used to construct a generalized subsolution/control requires finding a finite collection of affine subsolution/control pairs in $\mathbb{A}$ such that their minimum satisfies the appropriate terminal condition and takes a large value at $(0,0)$ [preferably 2γ, the optimal exponential decay rate]. We begin in this section with the simplest examples.

3.2.1 Example: Estimating $P\{X_n \in A\}$

Consider the setup in Section 2.1, and suppose that one wishes to estimate $P\{X_n\in A\}$ for some Borel set $A\subset\mathbb{R}^d$. Throughout this section we assume

$$\inf_{\beta\in A^\circ} L(\beta) = \inf_{\beta\in\bar{A}} L(\beta) \in (0,\infty),$$

where $A^\circ, \bar{A}$ denote the interior and the closure of A, respectively. It follows that the exponential decay rate is

$$\gamma = \inf_{\beta\in A} L(\beta) = \inf_{\beta\in A^\circ} L(\beta) = \inf_{\beta\in\bar{A}} L(\beta).$$

Let $\beta^*\in\bar{A}$ denote a minimizer of L over $\bar{A}$. Let α^* satisfy

$$L(\beta^*) = \sup_{\alpha\in\mathbb{R}^d} [\langle\alpha,\beta^*\rangle - H(\alpha)] = \langle\alpha^*,\beta^*\rangle - H(\alpha^*).$$

α^* is called the convex conjugate of β^*.

Case 1. Consider the simplest case where A is a convex set. Thanks to the convexity of $\bar{A}$, the vector $-\alpha^*$ defines an outward normal of $\bar{A}$. It follows that

$$A \subset \{x \in \mathbb{R}^d : \langle x, \alpha^* \rangle \geq \langle \beta^*, \alpha^* \rangle\}. \tag{3.1}$$

Consider the element of $\mathbb{A}$ with $\bar{\alpha} = \alpha^*$ and $\bar{c} = 2\langle \beta^*, \alpha^* \rangle$, i.e.,

$$\bar{W}(x,t) = -2\langle \alpha^*, x \rangle + 2\langle \beta^*, \alpha^* \rangle - 2(1-t)H(\alpha^*).$$

Thanks to the inequality (3.1), for each $x \in A$

$$\bar{W}(x,1) = -2\langle \alpha^*, x \rangle + 2\langle \beta^*, \alpha^* \rangle \leq 0.$$

Therefore, when A is convex, $(\bar{W}, \alpha^*)$ provides a simple subsolution/control pair that satisfies the terminal condition. The value at $(0,0)$ is

$$\bar{W}(0,0) = 2\langle \beta^*, \alpha^* \rangle - 2H(\alpha^*) = 2L(\beta^*) = 2\gamma,$$

the largest possible value. Note that the analysis holds if we replace the convexity assumption on A by the assumption that (3.1) holds.

Case 2. More generally, suppose that for some $K \in \mathbb{N}$,

$$A \subset \cup_{k=1}^{K} \{x : \langle x, \alpha_k \rangle \geq \langle \beta_k, \alpha_k \rangle\}$$

where β_k and α_k are convex conjugates, and that for each k $L(\beta_k) \geq \gamma$. A necessary and sufficient condition for these two assumptions to hold is that A should be contained in the union of a finite number of half-spaces, and that the infimum of L on each of these half spaces is at least γ. In this case β_k can be taken as the point on the kth half space that minimizes L, and we have $\gamma = L(\beta_k)$ for some k. Several of the numerical examples in Section 4 will fall into this category.

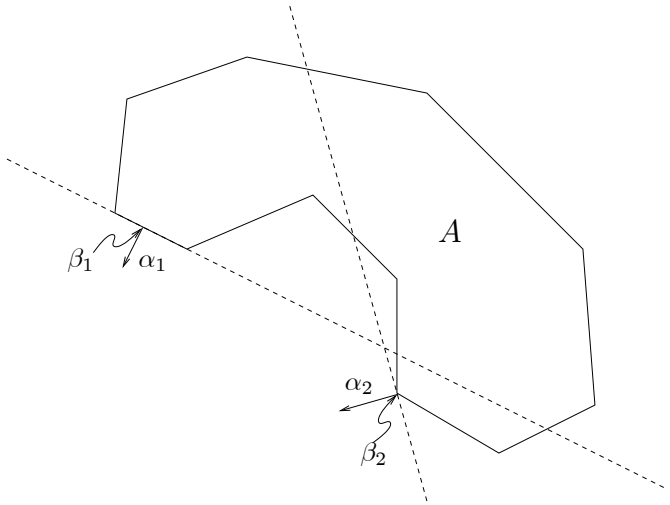

Figure 1. Example of a non-convex A with a two-piece subsolution.

Define an affine subsolution/control pair $(\bar{W}_k, \alpha_k)$ by

$$\bar{W}_k(x,t) \doteq -2\langle \alpha_k, x\rangle + 2\langle \alpha_k, \beta_k\rangle - 2(1-t)H(\alpha_k)$$

for each $k = 1, \dots, K$. Consider the pointwise minimum

$$\bar{W}(x,t) \doteq \wedge_{k=1}^{K} \bar{W}_k(x,t).$$

As we have pointed out, $\bar{W}$ defines a weak sense subsolution to the Isaacs equation. The terminal condition (2.7) is satisfied. Since for each $x \in A$ $\langle x, \alpha_k\rangle \geq \langle \beta_k, \alpha_k\rangle$ for some k, we have

$$\bar{W}(x,1) \leq \bar{W}_k(x,1) = -2\langle x, \alpha_k\rangle + 2\langle \beta_k, \alpha_k\rangle \leq 0.$$

Finally, we observe that

$$\bar{W}(0,0) = \wedge_{k=1}^{K} \bar{W}_k(0,0) = 2\wedge_{k=1}^{K} [\langle \beta_k, \alpha_k\rangle - 2H(\alpha_k)] = 2\wedge_{k=1}^{K} L(\beta_k) = 2\gamma.$$

The last equality holds since $L(\beta_k) \geq \gamma$ for each k and with equality for some k.

3.2.2 Example: Estimating $E\exp\{-nF(X_n)\}$

Consider the setup in Section 2.1, and suppose that one wishes to estimate $E\exp\{-nF(X_n)\}$ for some measurable function $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Let $\gamma \doteq \inf_{\beta\in\mathbb{R}^d} [F(\beta) + L(\beta)]$, and assume that

$$\lim_{n\to\infty} \frac{1}{n} \log E\exp\{-nF(X_n)\} = -\gamma.$$

This large deviation limit holds under various sets of regularity conditions on F [2]. In this case the terminal condition (2.7) is replaced by

$$\bar{W}(x, 1) \leq 2F(x)$$

for $x \in \mathbb{R}^d$ (see Remark 2.4).

The development here parallels the probability case as described in the previous section. For simplicity, we assume that there exists β^* that minimizes $L(\beta) + F(\beta)$ over $\beta \in \mathbb{R}^d$, and let α^* be its convex conjugate. To avoid technicalities, we also assume that L is differentiable at β^*, and thus

$$\alpha^* = DL(\beta^*).$$

Case 1. We first consider the simplest case where F is convex. Consider an affine subsolution/control pair $(\bar{W}, \alpha^*)$ where

$$\bar{W}(x, t) \doteq -2\langle\alpha^*, x\rangle + 2[F(\beta^*) + \langle\alpha^*, \beta^*\rangle] - 2(1 - t)H(\alpha^*).$$

Since β^* is a minimizer of $L(x) + F(x)$, we have

$$0 \in \partial(L + F)(\beta^*),$$

where ∂ denotes the set of subdifferentials. Therefore

$$-\alpha^* = -DL(\beta^*) \in \partial F(\beta^*).$$

It follows that the affine function $\bar{W}(x, 1)$ is a supporting hyperplane to $2F$ at β^*, and hence

$$\bar{W}(x, 1) \leq 2F(x)$$

for every x, i.e., the terminal condition holds. Also observe that

$$\bar{W}(0, 0) = 2F(\beta^*) + 2\langle\alpha^*, \beta^*\rangle - 2H(\alpha^*) = 2F(\beta^*) + 2L(\beta^*) = 2\gamma.$$

CASE 2. Next suppose that F is no longer convex. If there exists a convex function G such that $G \leq F$, $G(\beta^*) = F(\beta^*)$, and

$$\inf_{\beta\in\mathbb{R}^d} [L(\beta) + G(\beta)] = \inf_{\beta\in\mathbb{R}^d} [L(\beta) + F(\beta)] = \gamma,$$

then we reduce to the previous case. More generally, suppose there exist convex functions $G_k, k = 1, \ldots, K$, such that

$$\wedge_{k=1}^K G_k \leq F, \tag{3.2}$$

and for each k,

$$\inf_{\beta\in\mathbb{R}^d} [L(\beta) + G_k(\beta)] \geq \inf_{\beta\in\mathbb{R}^d} [L(\beta) + F(\beta)]\,. \tag{3.3}$$

If each G_k is bounded from below and lower semicontinuous then a minimizer β_k of $L(\beta) + G_k(\beta)$ will exist, and we can define the weak sense subsolution

$$\bar{W}(x,t) \doteq \wedge_{k=1}^K \bar{W}_k(x,t),$$

where $(\bar{W}_k, \alpha_k)$ is the affine subsolution/control pair with

$$\bar{W}_k(x,t) \doteq -2\langle \alpha_k, x\rangle + 2[G_k(\beta_k) + \langle \alpha_k, \beta_k\rangle] - 2(1-t)H(\alpha_k).$$

The same argument as in Case 1 shows that the terminal condition $\bar{W}(x,1) \leq 2F(x)$ is satisfied, and we have

$$\bar{W}(0,0) = \wedge_{k=1}^K [L(\beta_k) + G(\beta_k)] \geq 2\gamma.$$

Actually, the equality holds, since (3.2) implies (3.3) must hold as an equality for some k.

3.3 Mollification

As discussed previously, once a weak sense subsolution is identified as the pointwise minimum of a collection of affine subsolution/control pairs, mollification is used to produce generalized subsolution/controls.

Let $(\bar{W}_k, \alpha_k) \in \mathbb{A}, k = 1, 2 \ldots, K$, be affine subsolution/control pairs. We use a standard numerical approximation which we call *exponential weighting* for $\bar{W}(x,t) = \wedge_{k=1}^K \bar{W}_k(x,t)$. Let δ be a small positive number, and define

$$\bar{W}^\delta(x,t) \doteq -\delta \log\left(\sum_{k=1}^K e^{-\frac{1}{\delta}\bar{W}_k(x,t)}\right).$$

For $1 \le i \le K$, let

$$\rho_i^\delta(x,t) \doteq \frac{e^{-\frac{1}{\delta}\bar{W}_i(x,t)}}{\sum_{k=1}^K e^{-\frac{1}{\delta}\bar{W}_k(x,t)}}.$$

Then one can easily verify

$$D\bar{W}^\delta(x,t) = \sum_{k=1}^K \rho_k^\delta(x,t) D\bar{W}_k(x,t)$$

and

$$\bar{W}_t^\delta(x,t) = \sum_{k=1}^K \rho_k^\delta(x,t) \left(\bar{W}_k\right)_t (x,t),$$

and so $\bar{W}^\delta$ takes the form prescribed for a generalized subsolution/control with $\bar{\alpha}_k(x,t) \equiv \alpha_k$. It is obvious that the three functions $s_k(x,t) = D\bar{W}_k(x,t) = -2\alpha_k$, $r_k(x,t) = \left(\bar{W}_k\right)_t (x,t) = 2H(\alpha_k)$ and $\bar{\alpha}_k(x,t)$ are uniformly bounded and Lipschitz continuous, and it is easy to check that the same is true with regard to $\rho_k^\delta(x,t)$ for each fixed $\delta > 0$. Therefore, $(\bar{W}^\delta, \rho_k^\delta, \bar{\alpha}_k)$ is a generalized subsolution/control.

For a fixed (x,t) let l satisfy $\bar{W}(x,t) = \bar{W}_l(x,t)$. Then

$$e^{-\frac{1}{\delta}\bar{W}_k(x,t)} \le e^{-\frac{1}{\delta}\bar{W}_l(x,t)}$$

for $k = 1, \ldots, K$. Therefore

$$e^{-\frac{1}{\delta}\bar{W}_l(x,t)} \le \sum_{k=1}^K e^{-\frac{1}{\delta}\bar{W}_k(x,t)} \le K e^{-\frac{1}{\delta}\bar{W}_l(x,t)},$$

which implies

$$\bar{W}_l(x,t) \ge \bar{W}^\delta(x,t) \ge \bar{W}_l(x,t) - \delta \log K.$$

Since l achieves the minimum, for all (x,t)

$$\bar{W}(x,t) \ge \bar{W}^\delta(x,t) \ge \bar{W}(x,t) - \delta \log K.$$

Thus if $\bar{W}$ satisfies a given terminal condition then so will $\bar{W}^\delta$, though $\bar{W}^\delta(0,0) \ge \bar{W}(0,0) - \delta \log K$ implies there may be a small reduction of the value at $(0,0)$.

It is important to observe is that, as equation (2.5) indicates, the subsolution $\bar{W}^\delta$ itself does not play any explicit role in the computation of the change of measure and the algorithm is completely determined by $(\rho_k^\delta, \bar{\alpha}_k)$. However, the function $\bar{W}^\delta$ characterizes the performance of the corresponding importance sampling algorithm through results such as Theorem 2.2.

Remark 3.2 Recall that mollification will possibly results in a small reduction in the value $\bar{W}(0,0)$. The inequality $\bar{W}^\delta \leq \bar{W}$ suggests one could perhaps satisfy the terminal condition by considering $\bar{W}^\delta + c$ instead of $\bar{W}^\delta$ for some $c > 0$, which would reduce the loss at $(0,0)$. However, this will not eliminate the gap in all circumstances. It is worth noting that one can construct a sequence of schemes indexed by n which achieves the theoretical bound on performance if one chooses $\delta \to 0$ as $n \to \infty$ in an appropriate way. We will not pursue this issue here, since our computational experience suggests it is not needed to obtain good performance. However, the interested reader can consult [3] for the precise statement and further details in the context of stochastic networks.

Remark 3.3 Exponential weighting is not the only way to achieve mollification. For example, one can mollify $\bar{W}(x,t) = \wedge_{k=1}^K \bar{W}_k(x,t)$ by integration against a smooth convolution kernel, for which a standard choice is

$$\eta(x) \doteq \begin{cases} C \exp\{1/(\|x\|^2 - 1)\}, & \text{if } \|x\| < 1, \\ 0 \qquad\qquad\quad , & \text{if } \|x\| \geq 1, \end{cases}$$

where C is the normalizing constant so that the integral of η over $\mathbb{R}^d$ is one [7, Section 7.2]. However, we do not recommend this method since in this case the weights $\{\rho_k^\delta(x,t)\}$ involve integrations that can be computationally demanding. In contrast, the weights are very easy to compute when using the exponential weighting mollification. In fact, for all the numerical examples in Section 4 where mollification is needed, if one mollifies by integration against the convolution kernel η, the resulting importance sampling schemes will yield very similar estimates and standard errors (for the same sample size) as those based on the exponential weighting mollification, even though the latter is much faster.

3.4 Discussion

The construction of generalized subsolution/controls can be extended in many ways and to many other situations. Depending on the problem at hand, the Isaacs equation, the terminal condition, and the boundary conditions (if any) may take different forms (see Section 2.5). For example, when computing escape probabilities with a stochastic network one is particularly interested in combining subsolutions so as to satisfy appropriate boundary conditions [3]. In other problems, one way wish to expand $\mathbb{A}$ so that it also includes $(\bar{W}, \bar{\alpha})$ where $\bar{W}$ is a strict subsolution to the Isaacs equation, or even a non-affine subsolution of some specific form. However,

the basic structure for constructing an importance sampling scheme remains the same: We identify a class of subsolution/control pairs which correspond to changes of measures of simple form, and use these pairs as the building blocks for generalized subsolution/controls.

4 Examples of importance sampling algorithms

In this section we give examples of importance sampling algorithms based on subsolutions. As noted in the Introduction, some of these examples are not covered by the theoretical framework in the companion paper [6]. The role of these examples is to demonstrate the broad applicability of subsolutions in importance sampling. Unless specified otherwise, the importance sampling algorithm based on a generalized subsolution/control will follow the description in Section 2.3 with the new distribution determined by equation (2.5).

4.1 Example: Estimating $P\{X_n \in A\}$ for convex A

Assume that $\{Y_1, Y_2, \ldots\}$ is a sequence of iid 2-dimensional $N(0, I_2)$ random variables, where I_2 is the 2×2 identity matrix. Let

$$X_n = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

and consider the estimation of $P\{X_n \in A\}$ for a convex set A of the form

$$A = \left\{ x \in \mathbb{R}^2 : (x-a)^2 + y^2 \le r^2 \right\},$$

with $0 < r < a$. Clearly $\{X_n \in A\}$ is a rare event for large n.

For this model, $H(\alpha) = \|\alpha\|^2/2$ and $L(\beta) = \|\beta\|^2/2$. Cramér's Theorem yields

$$\lim_{n\to\infty} \frac{1}{n} \log P\{X_n \in A\} = -\inf_{\beta \in A} L(\beta) = -\frac{1}{2}(a-r)^2 \doteq -\gamma,$$

with the minimizing $\beta^* = (a-r, 0)$. The convex conjugate of β^* is just $\alpha^* = (a-r, 0)$.

As discussed in Section 3.2.1, the affine subsolution/control pair $(\bar{W}, \alpha^*)$

$$\bar{W}(x,t) = -2\langle \alpha^*, x \rangle + 2\langle \beta^*, \alpha^* \rangle - 2(1-t)H(\alpha^*)$$

satisfies the terminal condition (2.7) and $\bar{W}(0,0) = 2L(\beta^*) = 2\gamma$. By Theorem 2.2 the importance sampling algorithm based on $(\bar{W}, \alpha^*)$ is asymptotically optimal.

In this example, $(\bar{W}, \alpha^*)$ induces an importance sampling scheme that takes the simplest possible form, and indeed one that is well known in the importance sampling literature. Let μ be the underlying distribution $N(0, I_2)$. Then the $\{\bar{Y}_i\}$ are iid with distribution

$$\nu(dy) = e^{\langle \alpha^*, y \rangle - H(\alpha^*)} \mu(dy) = \frac{1}{2\pi} e^{-\|y-\alpha^*\|^2/2} dy,$$

which is the distribution of $N(\alpha^*, I_2)$.

The table below gives numerical results. We take $a = 2$, $r = 1$, and run simulations for the cases $n = 25, 50, 100$. Each estimate consists of 20,000 replications. The theoretical value (which is available from the standard statistics software S-plus) is presented for comparison. The standard error is also a numerical estimate, and C.I. stands for "confidence interval," though this is only formal since we make no assertion regarding normality of errors.

	$n = 25$	$n = 50$	$n = 100$
Theoretical value	1.99×10^{-7}	5.39×10^{-13}	5.36×10^{-24}
Estimate	2.00×10^{-7}	5.48×10^{-13}	5.44×10^{-24}
Standard Error	0.04×10^{-7}	0.12×10^{-13}	0.14×10^{-24}
95% C.I.	$[1.92, 2.08] \times 10^{-7}$	$[5.24, 5.72] \times 10^{-13}$	$[5.16, 5.72] \times 10^{-24}$

Table 1. Estimating $P\{X_n \in A\}$ for convex A.

Remark 4.1 This algorithm based on $(\bar{W}, \alpha^*)$ coincides with the importance sampling based on what one might call the "standard heuristic," which states that the change of measure used in the analysis of the large deviation lower bound is a good choice for importance sampling. As demonstrated in [9, 10, 4, 5], the standard heuristic importance sampling is efficient only in very special situations.

4.2 Example: Estimating $P\{X_n \in A\}$ for non-convex A

In this section, we give numerical estimates of $P\{X_n \in A\}$ when $A \subset \mathbb{R}^d$ takes the form

$$A \subset \{x : \langle x, \alpha_1 \rangle \geq \langle \beta_1, \alpha_1 \rangle\} \cup \{x : \langle x, \alpha_2 \rangle \geq \langle \beta_2, \alpha_2 \rangle\}, \tag{4.1}$$

with $\alpha_k \in \mathbb{R}^d$ and $\beta_k \in \mathbb{R}^d$ convex conjugates for $k = 1, 2$.

As discussed in Section 3.2.1, one can construct affine subsolution/control pairs $(\bar{W}_k, \alpha_k)$ with

$$\bar{W}_k(x,t) \doteq -2\langle \alpha_k, x\rangle + 2\langle \alpha_k, \beta_k\rangle - 2(1-t)H(\alpha_k)$$

such that their minimum

$$\bar{W}(x,t) \doteq \bar{W}_1(x,t) \wedge \bar{W}_2(x,t)$$

is a (weak sense) subsolution which satisfies the terminal condition (2.7).

Using the exponential weighting mollification scheme presented in Section 3.3, for any choice of δ one produces a generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \alpha_k)$. The importance sampling algorithm is determined by $(\rho_k^\delta, \alpha_k)$.

We will present two numerical examples: one for one-dimensional iid normal random variables, the other for a two-dimensional finite state Markov chain. Both examples have already appeared [4, 5]. The difference is that in these papers the importance sampling algorithm was based on the *solution* of the Isaacs equation, and whence the state dependence of the change of measure was much more involved.

4.2.1 IID normal random variables

Assume that $\{Y_1, Y_2, \ldots\}$ is a sequence of iid $N(0,1)$ random variables, and

$$X_n \doteq \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Suppose we are interested in estimating $P\{X_n \in A\}$ for the non-convex set

$$A = (-\infty, a] \cup [b, \infty)$$

with $a < 0 < b$. One can write A in the form of (4.1) by taking $\alpha_1 = \beta_1 = a$ and $\alpha_2 = \beta_2 = b$.

The importance sampling algorithm based on a generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \alpha_k)$ is as follows. We simulate $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ recursively. Let $\bar{X}_0 = 0$ and

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{j} \bar{Y}_j.$$

By (2.5), the conditional distribution of $\bar{Y}_{j+1}$ given $\bar{X}_j = x$ is

$$\sum_{k=1}^{2} \rho_k^\delta(x, j/n) \frac{1}{\sqrt{2\pi}} e^{(y-\alpha_k)^2/2} dy.$$

For numerical experimentation, we take $a = -0.25$, $b = 0.2$, and run simulations for $n = 100, 200, 500$. The mollification parameter δ is set as 0.02. Each estimate consists of 20,000 simulations.

	$n = 100$	$n = 200$	$n = 500$
Theoretical value	2.90×10^{-2}	2.54×10^{-3}	3.88×10^{-6}
Estimate	2.87×10^{-2}	2.50×10^{-3}	3.92×10^{-6}
Standard Error	0.03×10^{-2}	0.04×10^{-3}	0.08×10^{-6}
95% C.I.	$[2.81, 2.93] \times 10^{-2}$	$[2.42, 2.58] \times 10^{-3}$	$[3.76, 4.08] \times 10^{-6}$

Table 2. $P\{X_n \in A\}$ with one-dimensional, non-convex A.

4.2.2 A finite state Markov chain

Consider a two-node tandem Jackson network with arrival rate λ and consecutive rates μ_1, μ_2. We assume the queueing system is stable, that is, $\lambda < \mu_1 \wedge \mu_2$. The system has finite buffers of size B_1 and B_2, respectively.

The embedded time-homogeneous discrete-time Markov chain is $Y = \{Y_i = (Y_i^1, Y_i^2), i \in \mathbb{N}_0\}$, representing the queue lengths of the nodes at the epochs of transitions in the network. This process has the finite state space $S \doteq \{(y_1, y_2) : y_i = 0, 1, \ldots, B_i, i = 1, 2\}$. It is assumed that the system is initially empty, i.e., $Y_0 = (Y_0^1, Y_0^2) = (0, 0)$. The transition probability matrix of Y is denoted by P.

We are interested in estimating a class of probabilities associated with buffer overflow. More precisely, define $g : S \to \{0, 1\}^2$ by

$$g(y) \doteq \left(1_{\{y_1 = B_1\}}, 1_{\{y_2 = B_2\}}\right)$$

for every $y = (y_1, y_2) \in S$. Let

$$X_n \doteq \frac{1}{n} \sum_{i=0}^{n-1} g(Y_i).$$

We wish to estimate $P\{X_n \in A\}$, where A takes the form

$$A = \{(x_1, x_2) : x_1 \geq \varepsilon_1 \text{ or } x_2 \geq \varepsilon_2\}$$

for some $0 \leq \varepsilon_1, \varepsilon_2 \leq 1$.

For each $k = 1, 2$, let $\beta_k \in \mathbb{R}^2$ be the minimizer of $L(\beta)$ over the half space

$$H_k \doteq \{(x_1, x_2) : x_k \geq \varepsilon_k\}.$$

If $\alpha_k \in \mathbb{R}^2$ is the convex conjugate of β_k, then we can write A in the form of (4.1), that is,

$$A = \cup_{k=1}^2 H_k = \cup_{k=1}^2 \{x : \langle x, \alpha_k \rangle \geq \langle \beta_k, \alpha_k \rangle\}.$$

The importance sampling algorithm based on a generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \alpha_k)$ is as follows. We simulate $\{\bar{Y}_0, \bar{Y}_1, \ldots\}$ recursively, with initial state $\bar{Y}_0 = (0, 0)$. Let

$$\bar{X}_j = \frac{1}{n} \sum_{i=0}^{j} g(\bar{Y}_j).$$

Thanks to (2.5), the conditional distribution of $\bar{Y}_{j+1}$, given $\bar{X}_j = x$ and $\bar{Y}_j = y$, is the mixture

$$\sum_{k=1}^{2} \rho_k^\delta(x, j/n) P_k(y, \cdot),$$

where P_k is a transition probability matrix defined by (2.1) and (2.2):

$$P_k(y, z) = e^{\langle \alpha_k, g(z) \rangle - H(\alpha_k)} \frac{r(z; \alpha_k)}{r(y; \alpha_k)} P(y, z), \quad y, z \in S.$$

In the numerical simulation we take $B_1 = B_2 = 6$, and $\lambda = 0.2, \mu_1 = \mu_2 = 0.4$, and set $\varepsilon_1 = 0.3$, $\varepsilon_2 = 0.4$. The mollification parameter is chosen as $\delta = 0.1$. We run simulations for $n = 50, 80, 110$, and each estimate consists of 20,000 replications. The theoretical values are obtained using a recursive algorithm and presented for comparison.

	$n = 50$	$n = 80$	$n = 110$
Theoretical p_n	5.15×10^{-9}	3.47×10^{-12}	1.83×10^{-15}
Estimate	5.05×10^{-9}	3.39×10^{-12}	1.87×10^{-15}
Standard Error	0.21×10^{-9}	0.13×10^{-12}	0.08×10^{-15}
95% C.I.	$[4.63, 5.47] \times 10^{-9}$	$[3.13, 3.65] \times 10^{-12}$	$[1.71, 2.03] \times 10^{-15}$

Table 3. $P\{X_n \in A\}$ with two-dimensional, non-convex A.

4.3 Example: Estimating an expectation $E \exp\{-nF(X_n)\}$

Although the estimation of $E \exp\{-nF(X_n)\}$ is a generalization of the problem of estimating probabilities, as we saw in Section 3.2.2, the principle of constructing generalized subsolution/control is unchanged. Below we give a simple numerical experiment for illustrative purposes.

Consider a sequence of iid $N(0,1)$ random variables $\{Y_1, Y_2, \ldots\}$ and let

$$X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

We wish to estimate $E\exp\{-nF(X_n)\}$ where F is defined as follows:

$$F(x) \doteq \begin{cases} 0 & , \quad \text{if } \ x \le -(1+a^{-1}), \\ a(x+1)+1 & , \quad \text{if } \ -(1+a^{-1}) < x \le -1, \\ 1 & , \quad \text{if } \ -1 < x \le 1, \\ -b(x-1)+1 & , \quad \text{if } \ 1 < x \le 1+b^{-1}, \\ 0 & , \quad \text{if } \ x > 1+b^{-1}, \end{cases}$$

and where a, b are positive constants. Under these assumptions, we have $H(\alpha) = \alpha^2/2$ and $L(\beta) = \beta^2/2$, and

$$\lim_{n\to\infty} \frac{1}{n} \log E\exp\{-nF(X_n)\} = -\inf_{\beta\in\mathbb{R}} [F(\beta) + L(\beta)] \doteq -\gamma.$$

Note that F can be written as $G_1 \wedge G_2 \wedge G_3$, with

$$G_1(x) = (ax + a + 1)^+, \quad G_2(x) = 1, \quad G_3(x) = (bx - b - 1)^-$$

all convex functions. For each $k = 1, 2, 3$, let β_k be the minimizer of $L(\beta) + G_k(\beta)$ over $\beta \in \mathbb{R}^d$, and α_k the convex conjugate point of β_k, whence $\alpha_k = \beta_k$. Then equations (3.2) and (3.3) hold, and one can follow the general construction detailed in Section 3.2.2.

Let $(\bar{W}_k, \alpha_k)$ be the subsolution/control pair

$$\bar{W}_k(x,t) \doteq -2\langle \alpha_k, x\rangle + 2[G_k(\beta_k) + \langle \alpha_k, \beta_k\rangle] - 2(1-t)H(\alpha_k).$$

Their minimum $\bar{W} = \bar{W}_1 \wedge \bar{W}_2 \wedge \bar{W}_3$ is a (weak sense) subsolution that satisfies the terminal condition $\bar{W}(x,1) \le 2F(x)$ [Remark 2.4]. Furthermore,

$$\bar{W}(0,0) = 2\inf_k [G_k(\beta_k) + L(\beta_k)] = 2\inf_\beta [F(\beta) + L(\beta)] = 2\gamma,$$

the optimal decay rate.

A generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \alpha_k)$ obtained as in Section 3.3 induces the following importance sampling scheme. Let $\bar{X}_0 = 0$, and

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{j} \bar{Y}_i.$$

The sequence $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ is simulated recursively so that the conditional distribution of $\bar{Y}_{j+1}$, given $\bar{X}_j = x$, is the mixture of normal distributions

$$\sum_{k=1}^{3} \rho_k^{\delta}(x, j/n) \frac{1}{\sqrt{2\pi}} e^{(y-\alpha_k)^2/2}.$$

In the numerical simulation, we take $a = 3/2$, $b = 4$, and it is easy to check that $\alpha_1 = \beta_1 = -3/2$, $\alpha_2 = \beta_2 = 0$, and $\alpha_3 = \beta_3 = 5/4$. The mollification parameter δ is set to 0.1. We run simulations for $n = 10, 20, 30$, with 20,000 simulations for each estimate.

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	1.03×10^{-4}	1.87×10^{-8}	5.63×10^{-12}
Estimate	1.02×10^{-4}	1.86×10^{-8}	5.73×10^{-12}
Standard Error	0.01×10^{-4}	0.03×10^{-8}	0.09×10^{-12}
95% C.I.	$[1.00, 1.04] \times 10^{-4}$	$[1.80, 1.92] \times 10^{-8}$	$[5.58, 5.91] \times 10^{-12}$

Table 4. $E \exp\{-nF(X_n)\}$ for one-dimensional, non-convex F.

4.4 Example: Level crossing

In this section we consider importance sampling estimates for level crossing probabilities. To illustrate the main idea, we specialize to the following setup. Let $\{Y_1, Y_2, \ldots\}$ be a sequence of iid random vectors taking values in $\mathbb{R}^d$ with common distribution μ, and $A \subset \mathbb{R}^d$ a Borel set. Define the partial sum

$$S_n = \sum_{i=1}^{n} Y_i$$

with $S_0 = 0$, and for every real number $z > 0$ let

$$T_z = \inf \left\{ n \geq 0 : \frac{1}{z} S_n \in A \right\}.$$

Under certain conditions, $\{T_z < \infty\}$ is a rare event for large z and its probability will decay exponentially in the sense that

$$\lim_{z \to \infty} \frac{1}{z} \log P\{T_z < \infty\} = -\gamma$$

for some $\gamma > 0$. The simplest example is when $\{Y_i\}$ are iid, one dimensional random variables with a negative expectation and $A = (1, \infty)$. Naturally, the question is how to estimate $P\{T_z < \infty\}$ for large z.

The theoretical framework presented in [6] does *not* cover this case – for example the time horizon is now infinite (see Remark 4.2 for more information). However, the use of subsolutions still carries over and leads to simple and efficient importance sampling algorithms, and we will outline their use in the next few paragraphs.

Let H be the log-moment generating function for Y_1, and let L be the Legendre transform of H. It is not hard to argue that the Isaacs equation associated with level crossing problems is of the same form as (2.3), except that there is no time dependence. In other words, the Isaacs equation is

$$\sup_{\alpha\in\mathbb{R}^d}\inf_{\beta\in\mathbb{R}^d}\mathbb{H}(DW;\alpha,\beta)=0 \tag{4.2}$$

and with boundary condition $W(x)=0$ for $x\in A$. Here $W:\mathbb{R}^d\to\mathbb{R}$ and DW is its gradient, and as before

$$\mathbb{H}(s;\alpha,\beta)=\langle s,\beta\rangle+L(\beta)+\langle\alpha,\beta\rangle-H(\alpha)$$

for $s,\alpha,\beta\in\mathbb{R}^d$. A generalized subsolution/control $(\bar{W},\rho_k,\bar{\alpha}_k)$ is defined in a completely analogous fashion to Definition 2.1.

For a given generalized subsolution/control $(\bar{W},\rho_k,\bar{\alpha}_k)$, the corresponding importance sampling algorithm is as follows. Fix the parameter z. Let $\bar{X}_0=0$. Given $\bar{X}_j=x$, we simulate $\bar{Y}_{j+1}$ under the distribution

$$\sum_{k=1}^{K}\rho_k(x)P(dy;\bar{\alpha}_k(x)),$$

where $P(dy;\alpha)$ is the exponential twist

$$P(dy;\alpha)=e^{\langle\alpha,z\rangle-H(\alpha)}\mu(dz)$$

and with μ the distribution of Y_1. Finally, we update the dynamics and let

$$\bar{X}_{j+1}=\bar{X}_j+\frac{1}{z}\bar{Y}_{j+1}.$$

The simulation will be stopped at time

$$\bar{T}_z\doteq\inf\left\{n\geq 0:\bar{X}_n\in A\right\}=\inf\left\{n\geq 0:\frac{1}{z}\bar{S}_n\in A\right\},$$

and one forms the importance sampling estimator

$$Z^z\doteq 1_{\{\bar{T}_z<\infty\}}\cdot\prod_{j=0}^{\bar{T}_z-1}\left[\sum_{k=1}^{K}\rho_k(\bar{X}_j)\cdot e^{\left\langle\bar{\alpha}_k(\bar{X}_j),\bar{Y}_{j+1}\right\rangle-H\left(\bar{\alpha}_k(\bar{X}_j)\right)}\right]^{-1}.$$

The performance of this importance sampling algorithm can be characterized by a result analogous to Theorem 2.2, except that the terminal condition (2.7) should be replaced by the boundary condition

$$\bar{W}(x) \leq 0 \tag{4.3}$$

for all $x \in A$. Therefore, the goal is to construct a generalized subsolution/control $(\bar{W}, \rho_k, \alpha_k)$ of a simple form that satisfies the boundary condition (4.3) and such that $\bar{W}(0)$ is as close as possible to the optimal decay rate 2γ.

The construction follows the same path, that is, one first identifies a class of affine subsolution/control pairs and then builds a generalized subsolution/control by mollifying the minimum of such affine subsolution/control pairs as was done Section 3.3. The class of affine subsolution/control pairs that serve as building block, again denoted by $\mathbb{A}$, takes a different form in this setting:

$$\mathbb{A} \doteq \left\{ (\bar{W}, \bar{\alpha}) : \bar{W}(x) = -2\langle \bar{\alpha}, x \rangle + \bar{c},\ \bar{\alpha} \in \mathbb{R}^d, H(\bar{\alpha}) \leq 0, \bar{c} \in \mathbb{R} \right\}.$$

It is not difficult to check that every $(\bar{W}, \bar{\alpha}) \in \mathbb{A}$ is indeed is subsolution/control pair, since

$$\begin{aligned} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}; \bar{\alpha}, \beta) &= \inf_{\beta \in \mathbb{R}^d} \left[\langle -2\bar{\alpha}, \beta \rangle + L(\beta) + \langle \bar{\alpha}, \beta \rangle - H(\bar{\alpha}) \right] \\ &= \inf_{\beta \in \mathbb{R}^d} \left[\langle -\bar{\alpha}, \beta \rangle + L(\beta) \right] - H(\bar{\alpha}) \\ &= -2H(\bar{\alpha}) \\ &\geq 0. \end{aligned}$$

Analogous to the discussion in Section 3.2, suppose, for example, that there exists $K \in \mathbb{N}$ so that

$$A \subset \cup_{k=1}^{K} \left\{ x : \langle x, \bar{\alpha}_k \rangle \geq \bar{c}_k \right\}$$

for some $\bar{c}_k \in \mathbb{R}$ and $\bar{\alpha}_k \in \mathbb{R}^d$ such that $H(\bar{\alpha}_k) \leq 0$. Then for each k, one can define an affine subsolution/control pair $(\bar{W}_k, \bar{\alpha}_k) \in \mathbb{A}$ with

$$\bar{W}_k(x) \doteq -2\langle \bar{\alpha}_k, x \rangle + 2\bar{c}_k.$$

The minimum of $\{\bar{W}_k\}$ is a weak sense subsolution to the Isaacs equation (4.2) and it satisfies the boundary condition (4.3). One then mollifies in order to obtain a generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \bar{\alpha}_k)$.

Remark 4.2 In the theoretical analysis one needs to "bound" the infinite time horizon in a certain way – essentially to justify an approximation by a finite time problem – and then apply a verification argument similar the one used in [6]. More details can be found in [3], where analysis of this type is carried out for the problem of estimating buffer overflow probabilities in queueing networks.

We next present two examples. The first example is a one dimensional level crossing problem, which has been studied extensively [12, 11, 1], and we will see how the subsolution approach leads to the commonly used change of measure. The second example was studied in [10], where it was used as a counterexample to illustrate the danger of blindly following the standard heuristic approach to importance sampling.

For each example the numerical experiment considers exponential random variables. There are two reasons for choosing the exponential distribution. One is that an assumption we used very often to facilitate the analysis (e.g., in [4, 5, 6]) is that the log moment generating function H is finite everywhere. This is not true for exponential distributions, and as we will see in fact it is not necessary. The second reason is that for exponential distributions the level crossing probabilities can be explicitly computed, and these theoretical values can be used for comparison to indicate the accuracy of the importance sampling estimates.

4.4.1 One dimensional level crossing

Assume that $\{Y_1, Y_2, \ldots\}$ are iid random variables with common distribution μ and that $E[Y_i] < 0$. Let $S_n = Y_1 + \cdots + Y_n$ be the partial sum. Let $A = (1, \infty)$ and

$$T_z \doteq \inf\{n \geq 0 : S_n \in zA\} = \inf\{n \geq 0 : S_n > z\}.$$

Let H be the log moment generating function

$$H(\alpha) = \log \int_{\mathbb{R}} e^{\alpha y} \mu(dy),$$

and define $\bar{\alpha}$ to be the unique positive solution to $H(\alpha) = 0$. It is well known that, under mild conditions,

$$\lim_{z \to \infty} \frac{1}{z} \log P\{T_z < \infty\} = -\bar{\alpha},$$

or, $\gamma = \bar{\alpha}$. It obvious that

$$A \subset \{x : x \cdot \bar{\alpha} \geq \bar{\alpha}\},$$

and thus

$$\bar{W}(x) = -2\bar{\alpha}x + 2\bar{\alpha},$$

and $\bar{\alpha}$ are a subsolution/control pair which also satisfies the boundary condition $W(x) \leq 0$ for $x \in A$. Note that $\bar{W}(0) = 2\bar{\alpha} = 2\gamma$, whence the corresponding importance sampling algorithm is asymptotically optimal. This subsolution/control pair induces a change of measure that coincides with the classical choice, that is, the algorithm simulates iid $\{\bar{Y}_i\}$ with common distribution

$$\nu(dy) = e^{\bar{\alpha}y - H(\bar{\alpha})}\mu(dy).$$

For numerical experimentation, we consider the special case where for some constant $\theta > 0$, $Y_i + \theta$ is exponentially distributed with parameter λ. The assumption of $E[Y_i] < 0$ is equivalent to $\theta\lambda > 1$. A bit of algebra yields that $\bar{\alpha}$ is the unique positive root to the equation

$$0 = H(\alpha) = -\alpha\theta + \log\lambda - \log(\lambda - \alpha), \tag{4.4}$$

and that the simulation distribution is

$$\nu(dy) = (\lambda - \bar{\alpha})e^{-(\lambda - \bar{\alpha})(y+\theta)}dy.$$

Thus the $\{\bar{Y}_i + \theta\}$ are iid exponentially distributed with parameter $\lambda - \bar{\alpha}$. It is not difficult to show that $E\bar{Y}_i > 0$, and thus $\bar{T}_z$ is finite with probability one.

Below is a numerical result. We take $\lambda = 1$, $\theta = 2$, and run simulations for $m = 10, 20, 30$. Each estimate uses 20,000 simulations. The theoretical values are obtained by explicitly solving an associated integral equation (we omit the details). Indeed,

$$P\{T_z < \infty\} = \frac{\lambda - \bar{\alpha}}{\lambda}e^{-\bar{\alpha}z}. \tag{4.5}$$

The value of $\bar{\alpha}$ is obtained by numerically solving equation (4.4) using the bisection method, and $\bar{\alpha} \approx 0.80$.

	$m = 10$	$m = 20$	$m = 30$
Theoretical value	7.04×10^{-5}	2.44×10^{-8}	8.44×10^{-12}
Estimate	7.11×10^{-5}	2.44×10^{-8}	8.30×10^{-12}
Standard Error	0.07×10^{-5}	0.02×10^{-8}	0.08×10^{-12}
95% C.I.	$[6.97, 7.28] \times 10^{-5}$	$[2.40, 2.48] \times 10^{-8}$	$[8.14, 8.46] \times 10^{-12}$

Table 5. Estimating probability of level crossing for 1-dim random walk.

4.4.2 Two dimensional level crossing

Let $\{Y_n = (Y_n^1, Y_n^2), n \in \mathbb{N}\}$ be a sequence of iid random vectors with $E[Y_i^1] < 0$, $E[Y_i^2] < 0$, and common distribution μ. As before, denote the partial sum by

$$S_n = (S_n^1, S_n^2) \doteq \sum_{i=1}^{n} Y_i.$$

Let

$$A \doteq \{x = (x_1, x_2) : x_1 > 1 \text{ or } x_2 > 1\}.$$

It follows that

$$T_z = \inf\{n \geq 0 : S_n \in zA\} = \inf\left\{n \geq 0 : S_n^1 > z \text{ or } S_n^2 > z\right\}.$$

Let

$$H(\alpha) = \log \int_{\mathbb{R}^2} e^{\langle \alpha, y \rangle} \mu(dy)$$

and let $\bar{\alpha}_1 \doteq (\gamma_1, 0)$ and $\bar{\alpha}_2 \doteq (0, \gamma_2)$ where $\gamma_k > 0$ is the unique positive number such that $H(\bar{\alpha}_k) = 0$, $k = 1, 2$. Under mild conditions, we have

$$\lim_{z \to \infty} \frac{1}{z} \log P\{T_z < \infty\} = -\gamma \doteq -\gamma_1 \wedge \gamma_2.$$

We have

$$A \subset \{x : \langle x, \bar{\alpha}_1 \rangle \geq \gamma_1\} \cup \{x : \langle x, \bar{\alpha}_2 \rangle \geq \gamma_2\}.$$

For each $k = 1, 2$, an affine subsolution/control pair is $(\bar{W}_k, \bar{\alpha}_k)$ with

$$\bar{W}_k(x) \doteq -2\langle \bar{\alpha}_k, x \rangle + 2\gamma_k.$$

If $\bar{W} = \bar{W}_1 \wedge \bar{W}_2$, then $\bar{W}$ is a subsolution in a weak sense that satisfies the boundary condition (4.3). Also note that $\bar{W}(0) = 2(\gamma_1 \wedge \gamma_2) = 2\gamma$, the optimal decay rate. A generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \bar{\alpha}_k)$ is then be obtained by mollification, and the corresponding importance sampling algorithm is as follows. Let $\bar{S}_n = \bar{Y}_1 + \cdots + \bar{Y}_n$. We recursively simulate $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ such that given $\bar{S}_j/z = x$, the conditional distribution of $\bar{Y}_{j+1}$ is the mixture

$$\sum_{k=1}^{2} \rho_k^\delta(x) P(dy; \bar{\alpha}_k),$$

where

$$P(dy; \bar{\alpha}_k) = e^{\langle \bar{\alpha}_k, y \rangle - H(\bar{\alpha}_k)} \mu(dy).$$

We stop the simulation once the process $\{\bar{S}_n/z\}$ reaches the set A.

For the purpose of numerical experimentation, we consider the special case where for some constants $\theta_k > 0$, the distribution of $Y_1^k + \theta_k$ is exponential with parameter λ_k, $k = 1, 2$. Assume $\theta_k \lambda_k > 1$, or equivalently $E[Y_1^k] < 0$, for every $k = 1, 2$. We also assume that $\{Y_i^1\}$ and $\{Y_i^2\}$ are independent sequences. Under these conditions, for each k, $\gamma_k > 0$ is uniquely determined and satisfies the equation

$$0 = -\gamma_k \theta_k + \log \lambda_k - \log(\lambda_k - \gamma_k), \tag{4.6}$$

and

$$\begin{aligned} P(dy; \bar{\alpha}_1) &= (\lambda_1 - \gamma_1) e^{-(\lambda_1 - \gamma_1)(y_1 + \theta_1)} dy_1 \cdot \lambda_2 e^{-\lambda_2 (y_2 + \theta_2)} dy_2, \\ P(dy; \bar{\alpha}_2) &= \lambda_1 e^{-\lambda_1 (y_1 + \theta_1)} dy_1 \cdot (\lambda_2 - \gamma_2) e^{-(\lambda_2 - \gamma_2)(y_2 + \theta_2)} dy_2. \end{aligned}$$

Below is a numerical result. We take $\lambda_1 = \lambda_2 = 1$, and $\theta_1 = 2$, $\theta_2 = 3$. We run simulations for $m = 10, 20, 30$, and each estimate consists of 20,000 simulations. The theoretical values can be easily obtained from the one-dimensional formula (4.5). Again, the value of γ_k is obtained by numerically solving equation (4.6) using the bisection method, with $\gamma_1 \approx 0.80$ and $\gamma_2 \approx 0.86$. It is worth pointing out that the importance sampling estimate suggested by the standard heuristic, which is to simulate iid $\{\bar{Y}_i\}$ with distribution $P(dy; \bar{\alpha}_1)$, will have unbounded variance as z tends to infinity [10, Theorem 2(i)]. The mollification parameter δ is taken as 0.1.

	$m = 10$	$m = 20$	$m = 30$
Theoretical value	9.51×10^{-5}	2.88×10^{-8}	9.24×10^{-12}
Estimate	9.56×10^{-5}	2.87×10^{-8}	9.31×10^{-12}
Standard Error	0.10×10^{-5}	0.03×10^{-8}	0.09×10^{-12}
95% C.I.	$[9.36, 9.76] \times 10^{-5}$	$[2.81, 2.93] \times 10^{-8}$	$[9.13, 9.49] \times 10^{-12}$

Table 6. Probability of level crossing for 2-dim random walk.

4.5 Example: A path-dependent event

Let $\{Y_1, Y_2, \ldots\}$ be a sequence of iid random variables with common distribution μ and $E[Y_i] = 0$. As before, let H be the log-moment generating function and L its convex conjugate. Fix $n \in \mathbb{N}$, and for $1 \le i \le n$ define

$$X_i \doteq \frac{1}{n} \sum_{j=1}^{i} Y_j,$$

with $X_0 \doteq 0$. We are interested in estimating

$$E_n \doteq E\left[e^{-nF(X_n)} 1_{\{\max_{0\le i\le n} X_i \ge h\}}\right]$$

where $h > 0$ is a given constant. Let $\mathrm{AC}[0,1]$ denote the collection of all absolutely continuous functions on $[0,1]$. Assume that the large deviation limit

$$\lim_{n\to\infty} \frac{1}{n} \log E_n = -\gamma$$

holds, with γ the solution of the following variational problem:

$$\gamma = \inf\left\{\int_0^1 L(\dot\phi(t))\,dt + F(\phi(1)) : \phi \in \mathrm{AC}[0,1], \max_{0\le t\le 1} \phi(t) \ge h, \phi(0) = 0\right\}. \tag{4.7}$$

To write down the Isaacs equation associated with this estimation problem, we need to expand the state space to accommodate the path-dependence of the event. More precisely, the state process will be (X_i, B_i), where

$$B_i \doteq 1_{\left\{\max_{0\le j\le i} X_j \ge h\right\}}$$

is the indicator of whether or not the "barrier" h has been breached by step i. This problem can then be thought of as a combination of the level crossing problem of Section 4.4 and the finite time problem of Section 3.2. First consider the problem conditioned on $B_i = 1$. In this case the large deviation problem takes exactly the same form as in Section 3.2, and the subsolutions of interest are characterized by

$$\bar W_t(1,x,t) + \sup_{\alpha\in\mathbb{R}} \inf_{\beta\in\mathbb{R}} \mathbb{H}(D_x \bar W(1,x,t); \alpha, \beta) \ge 0$$

with $\mathbb{H}$ be as defined in (2.4), together with the terminal condition

$$\bar W(1,x,t) \le 2F(x). \tag{4.8}$$

An appropriate subsolution will give us a good importance sampling scheme for use at all times *after* the threshold is crossed. The question then is to identify the importance sampling scheme to use before the threshold is crossed. Let us suppose that as soon as the threshold is crossed we switch to the scheme associated with $\bar W(1,x,t)$, so that $\bar W(1,x,t)$ identifies an upper bound on the performance after this time. We are therefore back in the setting of the level crossing problem, though there is now an *exit cost*

$\bar{W}(1,h,t)$ depending on the (scaled) time that the barrier is crossed. Hence the subsolution for times prior to crossing the barrier is also

$$\bar{W}_t(0,x,t) + \sup_{\alpha\in\mathbb{R}} \inf_{\beta\in\mathbb{R}} \mathbb{H}(D_x\bar{W}(0,x,t);\alpha,\beta) \geq 0$$

(where the time derivative is needed because the exit cost depends on time), together with the boundary condition

$$\bar{W}(0,x,t) \leq \bar{W}(1,x,t) \tag{4.9}$$

for $x \geq 1, t \in [0,1]$.

Generalized subsolution/controls to these equations are defined as in Sections 4.4 and 3.2, and the corresponding importance sampling algorithm is as follows. Let $\bar{X}_0 = 0$ and $\bar{B}_0 = 0$. Given $\bar{X}_j = x$ and $\bar{B}_j = b$, we simulate $\bar{Y}_{j+1}$ under the distribution

$$\sum_{k=1}^{K} \rho_k(b,x,j/n) P(dy; \bar{\alpha}_k(b,x,j/n))$$

where $P(dy;\alpha)$ is the exponential twist

$$P(dy;\alpha) \doteq e^{\langle\alpha,y\rangle - H(\alpha)}\mu(dy).$$

Then we update the dynamics by

$$\bar{X}_{j+1} = \bar{X}_j + \frac{1}{n}\bar{Y}_{j+1}, \quad \bar{B}_{j+1} = 1_{\{\max_{0\leq i\leq j+1}\bar{X}_i \geq h\}}.$$

The performance of this importance sampling algorithm can be characterized by an analogous result to Theorem 2.2. In particular, if V^n is the second moment of a single sample of the importance sampling estimator corresponding to $\bar{W}$, then

$$\liminf_{n\to\infty} -\frac{1}{n}\log V^n \geq \bar{W}(0,0,0).$$

Therefore, the goal is to find a structurally simple generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$ satisfying (4.9) and (4.8) with the value of $\bar{W}(0,0,0)$ as large as possible, preferably equal to the optimal decay rate 2γ.

For illustration, we will consider the simple setting where

$$E_n = P\left\{\max_{0\leq i\leq n} X_i \geq h, X_n \leq l\right\},$$

for some constant $0 < l < h$. In other words, $F(x) = \infty$ for $x > l$ and 0 otherwise. As in the cases studied previously, a subsolution can be identified in terms of the solution to the large deviation variational problem. Using convexity and Jensen's inequality, this problem can be written in the form

$$\inf\left\{\rho_0 L\left(\frac{h}{\rho_0}\right) + \rho_1 L\left(\frac{l-h}{\rho_1}\right) : \rho_i \geq 0, i = 0, 1, \rho_0 + \rho_1 = 1\right\}.$$

Since the mean of μ is zero, $L(\beta) = 0$ if and only if $\beta = 0$, and thus the infimum is achieved at ρ_i^* with $\rho_i^* > 0$ for $i = 1, 2$. Let $\beta_0^* = h/\rho_0^*$ and $\beta_1^* = (l-h)/\rho_1^*$, and let α_0^* and α_1^* be the convex conjugates, respectively. We claim that $H(\alpha_0^*) = H(\alpha_1^*)$. Indeed, the necessary condition for a minimizer gives

$$L\left(\beta_0^*\right) - L'\left(\beta_0^*\right)\beta_0^* - L\left(\beta_1^*\right) + L'\left(\beta_1^*\right)\beta_1^* = 0$$

Using the characterization

$$\alpha_i^* = L'\left(\beta_i^*\right),$$

we have

$$H(\alpha_0^*) = L'\left(\beta_0^*\right)\beta_0^* - L\left(\beta_0^*\right) = L'\left(\beta_1^*\right)\beta_1^* - L\left(\beta_1^*\right) = H(\alpha_1^*).$$

Using that the interpretation of $\bar{W}(1, x, t)$ as the solution to the finite time problem with the given terminal condition, we know from Section 3.2 that a subsolution is given by

$$\bar{W}(1, x, t) = -2\alpha_1^* x + 2l\alpha_1^* - 2(1-t)H(\alpha_1^*).$$

As subsolution for the times prior to exceeding h we use the form

$$\bar{W}(0, x, t) = -2\alpha_0^* x + c_0 - 2(1-t)H(\alpha_0^*).$$

Since $H(\alpha_0^*) = H(\alpha_1^*)$, to satisfy the boundary condition we need $-2\alpha_0^* h + c_0 \leq -2\alpha_1^* h + 2\beta_1^*\alpha_1^*$, and to obtain the largest value for $\bar{W}(0,0,0)$ we take $c_0 = 2(\alpha_0^* - \alpha_1^*)h + 2l\alpha$, so that the two functions agree on $x = h$. It follows that

$$\begin{aligned}
\bar{W}(0,0,0) &= \left[\bar{W}(0,0,0) - \bar{W}(0,h,\rho_0^*)\right] + \left[\bar{W}(1,h,\rho_0^*) - \bar{W}(1,l,1)\right] \\
&\quad + \bar{W}(1,l,1) \\
&= 2\alpha_0^* h - \rho_0^* 2H(\alpha_0^*) + 2\alpha_1^*(l-h) - \rho_1^* 2H(\alpha_1^*) \\
&= 2\rho_0^*\left[\alpha_0^*\beta_0^* - 2H(\alpha_0^*)\right] + 2\rho_1^*\left[\alpha_1^*\beta_1^* - 2H(\alpha_1^*)\right] \\
&= 2\gamma.
\end{aligned}$$

In other words, the corresponding scheme is asymptotically optimal.

For a numerical example we take $Y_i \sim N(0,1)$. The corresponding importance sampling algorithm takes a very simple form. Let $\bar{X}_0 = 0$ and $\bar{B}_0 = 0$. If $\bar{B}_j = 0$, that is, the sample path maximum has not yet surpassed barrier h, we simulate $\bar{Y}_{j+1}$ under the distribution $N(2h - l, 1)$. If $\bar{B}_j = 1$, that is, the barrier h has already been reached, we simulate $\bar{Y}_{j+1}$ under the distribution $N(l - 2h, 1)$. Then we update the dynamics by

$$\bar{X}_{j+1} = \bar{X}_j + \frac{1}{n}\bar{Y}_{j+1}, \quad \bar{B}_{j+1} = 1_{\{\max_{0\le i\le j+1} \bar{X}_i \ge h\}}.$$

In the numerical experiment below, $h = 1$ and $l = 0.8$. Simulations were run for $n = 10, 20, 30$, and each estimate consists of 20,000 samples. What we call the "theoretical value" is an estimate based on 1 billion samples of the importance sampling scheme.

	$n = 10$	$n = 20$	$n = 30$
Theoretical value	1.68×10^{-5}	9.66×10^{-9}	6.09×10^{-12}
Estimate	1.74×10^{-5}	9.58×10^{-9}	6.26×10^{-12}
Standard Error	0.04×10^{-5}	0.27×10^{-9}	0.19×10^{-12}
95% C.I.	$[1.66, 1.82] \times 10^{-5}$	$[9.04, 10.12] \times 10^{-9}$	$[5.88, 6.64] \times 10^{-12}$

Table 7. Estimating a path-dependent probability.

4.6 Example: A mixed open-closed queueing network

Consider the mixed open and closed queueing network as shown in Figure 2.

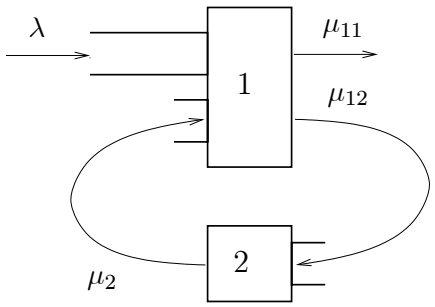

Figure 2. An open/closed queueing network.

The open jobs arrive at server 1 according to a Poisson process with rate λ. There is one closed job that circulates between server 1 and server 2, and it has pre-emptive priority over open jobs at server 1. All services rates are exponentially distributed. The service rates at server 1 are μ_{11} for open jobs and μ_{12} the closed job, and the service rate at server 2 is μ_2. Note that this system is equivalent to an $M/M/1$ queue with server breakdowns or vacations.

The state of the system is described by process $\{(Y_t, Z_t) : t \geq 0\}$, where Y_t and Z_t are the numbers of open jobs and closed jobs at server 1 at time t. We wish to estimate p_n, the probability that the number of open jobs reaches n before the system returns to state $(0, 0)$, given that the system starts in $(0, 0)$.

In general, the system can have multiple closed jobs. Even though we only consider the case of one closed job, the approach works for the general case, with the sole difference being that the computation of γ becomes more involved.

The following stability assumption will be in effect throughout this section:

$$\frac{\lambda}{\mu_{11}} + \frac{\mu_2}{\mu_2 + \mu_{12}} < 1. \tag{4.10}$$

Under this assumption, the system is stable, and p_n is a rare-event probability when n gets large.

The associated large deviation asymptotics can be characterized by an ODE. More precisely, let $y \in \{0, 1, \ldots, n\}$, $z \in \{0, 1\}$, and $V_n(y, z)$ the probability that the number of open jobs reaches n before the system returns to state $(0, 0)$, given that the system starts in (y, z) [whence $p_n = V_n(0, 0)$ by definition]. Given any $x \in [0, 1]$ and $z \in \{0, 1\}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log V_n(\lfloor nx \rfloor, z) = -v(x)$$

where v is a viscosity solution to an ordinary differential equation (ODE). To describe this ODE, we define the convex function

$$\ell(s) \doteq \begin{cases} s \log s - s + 1, & s \geq 0 \\ +\infty \quad , & s < 0 \end{cases},$$

and introduce the notation $\rho = (\rho_0, \rho_1), \hat{\Theta} = (\hat{\lambda}_0, \hat{\lambda}_1, \hat{\mu}_{11}, \hat{\mu}_{12}, \hat{\mu}_2)$. For $\beta \in \mathbb{R}$ define

$$L(\beta) \doteq \inf_{\rho, \hat{\Theta}} G(\rho, \hat{\Theta}), \tag{4.11}$$

where

$$G(\rho,\hat{\Theta}) \doteq \rho_0\left[\lambda\ell\left(\frac{\hat{\lambda}_0}{\lambda}\right)+\mu_{12}\ell\left(\frac{\hat{\mu}_{12}}{\mu_{12}}\right)\right] + \rho_1\left[\lambda\ell\left(\frac{\hat{\lambda}_1}{\lambda}\right)+\mu_{11}\ell\left(\frac{\hat{\mu}_{11}}{\mu_{11}}\right)+\mu_2\ell\left(\frac{\hat{\mu}_2}{\mu_2}\right)\right]$$

and with the infimum in (4.11) taken over all $(\rho,\hat{\Theta})$ such that

$$\rho_0\geq 0,\ \rho_1\geq 0,\ \rho_0+\rho_1=1,\ \rho_0\hat{\mu}_{12}=\rho_1\hat{\mu}_2,\ \beta=\rho_0\hat{\lambda}_0+\rho_1\hat{\lambda}_1-\rho_1\hat{\mu}_{11}. \quad (4.12)$$

One can show that function L is indeed convex and explicitly calculate the Legendre transform H of L. Indeed, we have (see the appendix for details)

$$H(\alpha)=\inf_q[H_0(\alpha,q)\vee H_1(\alpha,q)] \quad (4.13)$$

where

$$\begin{aligned} H_0(\alpha,q) &= \lambda(e^\alpha-1)+\mu_{12}(e^q-1),\\ H_1(\alpha,q) &= \lambda(e^\alpha-1)+\mu_{11}(e^{-\alpha}-1)+\mu_2(e^{-q}-1). \end{aligned}$$

Then v satisfies the ODE

$$0=\inf_{\beta\in\mathbb{R}}\left[v'(x)\cdot\beta+L(\beta)\right]=-H(-v'(x)),$$

with the boundary condition $v(1)=0$. Solving this ODE (again see the appendix for details), we obtain

$$v(x)=\gamma(1-x), \quad (4.14)$$

where

$$\gamma=-\log\frac{(\lambda+\mu_{11}+\mu_{12}+\mu_2)+\sqrt{(\lambda+\mu_{11}+\mu_{12}+\mu_2)^2-4\mu_{11}(\lambda+\mu_{12})}}{2\mu_{11}(1+\lambda^{-1}\mu_{12})}$$

is a positive number. In particular,

$$\lim_{n\to\infty}\frac{1}{n}\log p_n=-v(0)=-\gamma.$$

Moreover, the minimizing $\hat{\Theta}$ for (4.11) is

$$\Theta^*=(\lambda_0^*,\lambda_1^*,\mu_{11}^*,\mu_{12}^*,\mu_2^*)=(e^\gamma\lambda,e^\gamma\lambda,e^{-\gamma}\mu_{11},e^{q^*}\mu_{12},e^{-q^*}\mu_2), \quad (4.15)$$

where q^* is the minimizer in equation (4.13) with $\alpha = \gamma$, or

$$q^* = \log \frac{\lambda + \mu_{12} - \lambda e^{\gamma}}{\mu_{12}}. \tag{4.16}$$

Now let us consider the construction of importance sampling algorithms. We first describe the associated Isaacs equation. Let $\bar{\Theta} \doteq (\bar{\lambda}_0, \bar{\lambda}_1, \bar{\mu}_{11}, \bar{\mu}_{12}, \bar{\mu}_2)$. Define

$$\bar{L}(\beta, \bar{\Theta}) \doteq \inf_{\rho, \hat{\Theta}} \left[2G(\rho, \hat{\Theta}) - \bar{G}(\rho, \hat{\Theta}, \bar{\Theta}) \right] \tag{4.17}$$

where

$$\begin{aligned} \bar{G}(\rho, \hat{\Theta}, \bar{\Theta}) \doteq \rho_0 & \left[\bar{\lambda}_0 \ell \left(\frac{\hat{\lambda}_0}{\bar{\lambda}_0} \right) + \bar{\mu}_{12} \ell \left(\frac{\hat{\mu}_{12}}{\bar{\mu}_{12}} \right) \right] \\ & + \rho_1 \left[\bar{\lambda}_1 \ell \left(\frac{\hat{\lambda}_1}{\bar{\lambda}_1} \right) + \bar{\mu}_{11} \ell \left(\frac{\hat{\mu}_{11}}{\bar{\mu}_{11}} \right) + \bar{\mu}_2 \ell \left(\frac{\hat{\mu}_2}{\bar{\mu}_2} \right) \right] \end{aligned}$$

and the infimum in (4.17) is taken over all $(\rho, \hat{\Theta})$ satisfying the constraints (4.12). The Isaacs equation associated with importance sampling can then be written as

$$\sup_{\bar{\Theta}} \inf_{\beta} \left[W'(x) \cdot \beta + \bar{L}(\beta, \bar{\Theta}) \right] = 0,$$

with boundary condition $W(1) = 0$. Let $\bar{W} = 2v$. It is not difficult to show (see the appendix for details) that $(\bar{W}, \Theta^*)$ defines a affine subsolution/control pair to the Isaacs equation, and it satisfies the terminal condition $\bar{W}(1) = 2v(1) = 0$. Furthermore, $\bar{W}(0) = 2v(0) = 2\gamma$, the optimal decay rate.

The importance sampling algorithm corresponding to this affine subsolution/control pair $(\bar{W}, \Theta^*)$ is very simple. When $z = 1$, we simulate the system under the alternative probability measure such that the open job arrival rate is λ_1^* and the service rate for the closed job at server 1 is μ_{12}^*. When $z = 0$, the simulation distribution is such that the open job arrival rate is λ_0^* and the service rate for the open job is μ_{11}^* and the service rate for the closed job at server 2 is μ_2^*. Note that, for this special network, since $\lambda_0^* = \lambda_1^* \doteq \lambda^*$, the above change of measure is equivalent to simulation under the alternative rates $(\lambda^*, \mu_{11}^*, \mu_{12}^*, \mu_2^*)$.

In the numerical example below we take $\lambda = 1, \mu_{11} = 4, \mu_{12} = 2, \mu_2 = 0.5$. It is easy to check that the stability condition (4.10) holds. We run simulations for $n = 20, 40, 80$, and each estimate consists of 20,000 simulations. The theoretical value can be found in [8].

	$n = 20$	$n = 40$	$n = 80$
Theoretical value	3.91×10^{-8}	2.02×10^{-15}	5.40×10^{-30}
Estimate	3.93×10^{-8}	2.01×10^{-15}	5.45×10^{-30}
Standard Error	0.03×10^{-8}	0.02×10^{-15}	0.04×10^{-30}
95% C.I.	$[3.87, 3.99] \times 10^{-8}$	$[1.97, 2.05] \times 10^{-15}$	$[5.37, 5.53] \times 10^{-30}$

Table 8. Overflow probabilities of a mixed open-closed queueing network.

4.7 Example: A "universal" importance sampling scheme

For all the examples we have discussed, finitely many affine subsolution/control pairs are used to construct a generalized subsolution/control. In this section, we present an example where infinitely many subsolution/control pairs are used for this purpose. The corresponding importance sampling scheme has some interesting features, which are further discussed in Remark 4.3.

For illustration, we consider again the simple setting where $\{Y_1, Y_2, \ldots\}$ is a sequence of iid random variables taking values in $\mathbb{R}^d$ and let

$$X_n \doteq \frac{1}{n} \sum_{i=1}^{n} Y_i$$

with $X_0 \doteq 0$. We wish to estimate $P\{X_n \in A\}$ where $A \subset \mathbb{R}^d$ is a Borel set, and assume a large deviation limit holds, that is,

$$\lim_{n\to\infty} \frac{1}{n} \log P\{X_n \in A\} = -\inf_{\beta \in A} L(\beta) = -\gamma.$$

A new way to construct a subsolution is as follows. Consider the level set of the rate function L

$$\Theta_\gamma \doteq \left\{\beta \in \mathbb{R}^d : L(\beta) \geq \gamma\right\}.$$

It follows easily that Θ_γ is the complement of a convex set and that $A \subset \Theta_\gamma$. For each $\beta \in \partial\Theta_\gamma$, let $\alpha(\beta)$ be its convex conjugate (assuming its existence). Define $(\bar{W}_\beta, \alpha(\beta)) \in \mathbb{A}$ by

$$\bar{W}_\beta(x,t) \doteq -2\langle\alpha(\beta), x\rangle + 2\langle\alpha(\beta), \beta\rangle - 2(1-t)H[\alpha(\beta)].$$

Let

$$\bar{W}(x,t) \doteq \inf\left\{\bar{W}_\beta(x,t) : \beta \in \partial\Theta_\gamma\right\}.$$

Then $\bar{W}$ defines a (weak sense) subsolution to the Isaacs equation (2.3). Furthermore, thanks to the convexity of Θ_γ, we have

$$\bar{W}(x,1) = \inf\left\{-2\langle\alpha(\beta), x - \beta\rangle : \beta \in \partial\Theta_\gamma\right\} \leq 0$$

for $x \in \Theta_\gamma$. In particular, $\bar{W}$ satisfies the terminal condition (2.7) since $A \subset \Theta_\gamma$. In many cases (e.g., when L is finite), we have

$$\begin{aligned}\bar{W}(0,0) &= \inf\{2\langle\alpha(\beta),\beta\rangle - 2H[\alpha(\beta)] : \beta \in \partial\Theta_\gamma\} \\ &= \inf\{2L(\beta) : \beta \in \partial\Theta_\gamma\} \\ &= 2\gamma,\end{aligned}$$

the optimal decay rate.

Using Remark 2.2 and Theorem 2.2, if $\bar{W}$ were continuously differentiable then $(\bar{W}, \bar{\alpha})$ with $\alpha(x,t) = -D\bar{W}(x,t)/2$ would form a subsolution/control pair, and the corresponding importance sampling schemes would be asymptotically optimal performance. Since $\bar{W}$ is not continuously differentiable, one possibility is to resort to mollification. However, an alternative that is possible in some cases is to show that for each $\varepsilon > 0$ one can find a smooth function $\bar{W}^\varepsilon$ such that $(\bar{W}^\varepsilon, \bar{\alpha})$ form a subsolution/control pair and $\bar{W}^\varepsilon \to \bar{W}$ as $\varepsilon \to 0$. This approach is used in the following example.

To give a concrete example, we will work out the details for iid $N(0, I_d)$ sequence $\{Y_1, Y_2, \ldots\}$, where I_d denotes the identity matrix of dimension d. It follows that $H(\alpha) = \|\alpha\|^2/2$, $L(\beta) = \|\beta\|^2/2$, and whence

$$\Theta_\gamma = \left\{\beta \in \mathbb{R}^d : \|\beta\| \geq \sqrt{2\gamma}\right\}.$$

In this case $\alpha(\beta) = \beta$, and

$$\begin{aligned}\bar{W}(x,t) &\doteq \inf\{-2\langle\beta,x\rangle + 2\langle\beta,\beta\rangle - 2(1-t)H(\beta) : \|\beta\| \in \partial\Theta_\gamma\} \\ &= \inf\left\{-2\langle\beta,x\rangle + 2(1+t)\gamma : \|\beta\| = \sqrt{2\gamma}\right\} \\ &= -2\sqrt{2\gamma}\|x\| + 2(1+t)\gamma.\end{aligned}$$

It is not difficult to check by direct computation that $\bar{W}$ satisfies the Isaacs equation (2.3) except at $\{x = 0\}$, and $\bar{W}(x,1) \leq 0$ on Θ_γ. In particular, we have $\bar{W}(0,0) = 2\gamma$.

Even though $\bar{W}$ is not continuously differentiable at $\{x = 0\}$, it induces a control

$$\bar{\alpha}(x,t) \doteq -\frac{D\bar{W}(x,t)}{2} = \sqrt{2\gamma}\frac{x}{\|x\|}, \quad \text{if } x \neq 0.$$

For $x = 0$, we just define

$$\bar{\alpha}(x,0) \doteq \sqrt{2\gamma}\theta, \tag{4.18}$$

where θ is an arbitrarily fixed unit vector. We claim that the importance sampling scheme corresponding to this control $\bar{\alpha}$ is asymptotically optimal.

Indeed, consider the approximating sequence $\bar{W}^\varepsilon$ defined as follows. Let

$$\bar{W}^\varepsilon(x,t) \doteq -2\sqrt{2\gamma}\sqrt{\|x\|^2+\varepsilon} + 2(1+t)\gamma.$$

It is not difficult to check that, for every $\varepsilon > 0$, $\bar{W}^\varepsilon$ is continuously differentiable and $(\bar{W}^\varepsilon, \bar{\alpha})$ is a subsolution/control pair. The asymptotic optimality follows if one applies Theorem 2.2 to this subsolution/control pair, and observes that

$$\lim_{\varepsilon\to 0} W^\varepsilon(0,0) = \bar{W}(0,0) = 2\gamma,$$

the optimal decay rate.

For numerical experimentation, we take $d = 2$ and

$$A \doteq \left\{x = (x_1, x_2) : (x_1 + a)^2 + x_2^2 \geq R^2\right\}$$

for some constants $0 < a < R$. It follows that

$$\gamma = \inf_{\beta\in A} L(\beta) = \frac{1}{2}(R-a)^2,$$

with minimizer $\beta^* = (R-a, 0)$. Setting $\theta = (1,0)$ in equation (4.18), the importance sampling scheme is as follows. We recursively simulate $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ and let

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{j} \bar{Y}_i.$$

The conditional distribution of $\bar{Y}_{j+1}$ given $\bar{X}_j = x = (x_1, x_2)$ is

$$N\left(\left[\begin{array}{c} (R-a)x_1/\|x\| \\ (R-a)x_2/\|x\| \end{array}\right], I_2\right)$$

if $x \neq 0$, and

$$N\left(\left[\begin{array}{c} R-a \\ 0 \end{array}\right], I_2\right)$$

if $x = 0$.

For the table below we take $R = 0.5, a = 0.05$. We run simulations for $n = 40, 80, 120$, and each estimate consists of 20,000 simulations. The theoretical value can be obtained using standard software such as S-plus.

	$n = 40$	$n = 80$	$n = 120$
Theoretical value	8.49×10^{-3}	1.00×10^{-4}	1.40×10^{-6}
Estimate	8.38×10^{-3}	1.04×10^{-4}	1.50×10^{-6}
Standard Error	0.18×10^{-3}	0.04×10^{-4}	0.07×10^{-6}
95% C.I.	$[8.02, 8.74] \times 10^{-3}$	$[0.96, 1.12] \times 10^{-4}$	$[1.36, 1.64] \times 10^{-6}$

Table 9. A "universal" scheme

Remark 4.3 An important feature of this approach is that the construction of the importance sampling scheme does not need any information on the set A other than γ, the infimum of $L(\beta)$ over $\beta \in A$. It is for this reason that the scheme might be called "universal." The scheme has strengths and weaknesses. On one hand, the scheme is simple and applicable in more general settings. On the other hand, it is often the case that one knows some more detailed properties of target set A, which can be used to design more efficient schemes. Furthermore, there is a practical computational issue of obtaining $\bar{W}$ as the minimum of infinitely many subsolutions in the case of, say, sums of functionals of a Markov chain. However, here one may be willing to approximate Θ_γ by a finite number of points and then use exponential weighting for mollification.

4.8 Summary

We have shown that importance sampling schemes based on subsolutions can be applied in a wide variety of settings and deliver excellent performance. Besides being fast and accurate, the behavior of the schemes is very stable across a broad range of problem formulations. For example, in each setting and for each simulation we use the same number of samples (20,000) with remarkably similar performance. Moreover the asymptotic behavior of the schemes can be backed up by rigorous theoretical justification. Both of these properties stand in sharp contrast to the instability and poor performance exhibited by standard heuristic importance sampling schemes in many situations [10, 9, 4, 5].

5 Appendix. Computation for the mixed open-closed network example

Proof of Equation (4.13). We first define two functions. Let $\delta = (\delta_1, \delta_2) \in \mathbb{R}^2$, and define

$$L_0(\delta) \doteq \lambda \ell\left(\frac{\delta_1}{\lambda}\right) + \mu_{12}\ell\left(\frac{\delta_2}{\mu_{12}}\right).$$

Similarly, let $\eta = (\eta_1, \eta_2) \in \mathbb{R}^2$ and define

$$L_1(\eta) \doteq \inf\left\{\lambda\ell\left(\frac{\hat{\lambda}_1}{\lambda}\right) + \mu_{11}\ell\left(\frac{\hat{\mu}_{11}}{\mu_{11}}\right) + \mu_2\ell\left(\frac{\hat{\mu}_2}{\mu_2}\right)\right\}$$

where the infimum is taken over all $(\hat{\lambda}_1, \hat{\mu}_{11}, \hat{\mu}_2)$ such that

$$\hat{\lambda}_1 - \hat{\mu}_{11} = \eta_1, \quad -\mu_2 = \eta_2.$$

It is not difficult to show by direct computation that L_i is convex and its convex conjugate is H_i, for each $i = 0, 1$. Given $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$, let

$$Q(\theta) \doteq \inf \{\rho_0 L_0(\delta) + \rho_1 L_1(\eta) : \rho_0 \geq 0, \rho_1 \geq 0, \rho_0 + \rho_1 = 1, \rho_0 \delta + \rho_1 \eta = \theta\}.$$

Thanks to [2, Corollary D.4.3], G is a convex function whose convex conjugate is $H_0 \vee H_1$. Also, it is easy to see by definition that

$$L(\beta) = Q(\beta, 0).$$

Thus L is convex and its convex conjugate is H.

Proof of Equations (4.14) – (4.16). Consider the equation

$$H(\alpha) = \inf_q [H_0(\alpha, q) \vee H_1(\alpha, q)] = 0.$$

For every fixed α, $H_0(\alpha, q)$ is a strictly increasing function of q and $\lim_{q\to\infty} H_0(\alpha, q) = +\infty$. Similarly, for each fixed α, $H_1(\alpha, q)$ is a strictly decreasing function of q and $\lim_{q\to-\infty} H_1(\alpha, q) = +\infty$. It follows that, for each fixed α, there exists a unique $q = q(\alpha)$ such that

$$H_0(\alpha, q(\alpha)) = H_1(\alpha, q(\alpha)) = H(\alpha).$$

Therefore, solving $H(\alpha) = 0$ is equivalent to finding (α, q) such that

$$H_0(\alpha, q) = H_1(\alpha, q) = 0,$$

which yields the γ of (4.14) and the q^* of (4.16). Thus (γ, q^*) satisfies

$$H_0(\gamma, q^*) = H_1(\gamma, q^*) = 0. \tag{5.1}$$

(It turns out that $(0, 0)$ is also a solution, but it is elementary that this root does not characterize the relevant solution to the PDE.) The computation for Θ^* is straightforward and thus omitted. ■

Proof that $(\bar{W}, \Theta^*)$ is a subsolution/control pair. We only need to show that

$$\inf_\beta \left[-2\gamma\beta + \bar{L}(\beta, \Theta^*)\right] \geq 0. \tag{5.2}$$

However, analogous to the proof of equation (4.13), one can show that $L(\beta, \Theta^*)$ is convex with respect to β and its Legendre transform, denoted by $\bar{H}(\alpha)$, is

$$\bar{H}(\alpha) = \inf_q \left[\bar{H}_0(\alpha, q) \vee \bar{H}_1(\alpha, q)\right]$$

with

$$\begin{aligned}
\bar{H}_0(\alpha, q) &= \left(\frac{\lambda^2}{\lambda_0^*} e^{\alpha} - 2\lambda + \lambda_0^*\right) + \left(\frac{\mu_{12}^2}{\mu_{12}^*} e^{q} - 2\mu_{12} + \mu_{12}^*\right) \\
\bar{H}_1(\alpha, q) &= \left(\frac{\lambda^2}{\lambda_1^*} e^{\alpha} - 2\lambda + \lambda_1^*\right) + \left(\frac{\mu_{11}^2}{\mu_{11}^*} e^{-\alpha} - 2\mu_{11} + \mu_{11}^*\right) + \left(\frac{\mu_2^2}{\mu_2^*} e^{-q} - 2\mu_2 + \mu_2^*\right).
\end{aligned}$$

Then the inequality (5.2) reduces to $-\bar{H}(2\gamma) \geq 0$. Indeed, we claim that

$$\bar{H}(2\gamma) = 0. \tag{5.3}$$

To this end, note that

$$\bar{H}(2\gamma) = \inf_q \left[\bar{H}_0(2\gamma, q) \vee \bar{H}_1(2\gamma, q)\right].$$

However, by direction computation and equations (4.15), (5.1), we have

$$\bar{H}_0(2\gamma, 2q^*) = 2\lambda(e^{\gamma} - 1) + 2\mu_{12}(e^{q^*} - 1) = 2H_0(\gamma, q^*) = 0$$

and

$$\bar{H}_1(2\gamma, 2q^*) = 2\lambda(e^{\gamma} - 1) + 2\mu_{11}(e^{-\gamma} - 1) + 2\mu_2(e^{-q^*} - 1) = 2H_1(\gamma, q^*) = 0.$$

It is now very easy to argue that

$$\bar{H}(2\gamma) = \inf_q \left[\bar{H}_0(2\gamma, q) \vee H_1(2\gamma, q)\right] = \bar{H}_0(2\gamma, 2q^*) \vee H_1(2\gamma, 2q^*) = 0,$$

which completes the proof. ■

References

[1] S. Asmussen. *Ruin Probabilities*. World Scientific, Singapore, 2000.

[2] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.

[3] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Preprint*, 2005.

[4] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports.*, 76:481–508, 2004.

[5] P. Dupuis and H. Wang. Dynamic importance sampling for uniformly recurrent Markov chains. *Annals of Applied Probab.*, 15:1–38, 2005.

[6] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Convergence analysis. *Preprint*, 2005.

[7] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer–Verlag, Berlin, second edition, 1983.

[8] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47:585–600, 1999.

[9] P. Glasserman and S. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Modeling Comp. Simulation*, 4:22–42, 1995.

[10] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.*, 7:731–746, 1997.

[11] T. Lehtonen and H. Nyrhinen. Simulating level-crossing probabilities by importance sampling. *Adv. Appl. Prob.*, pages 858–874, 1992.

[12] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4:673–684, 1976.